ROBUST AUDIO HASHING FOR AUDIO IDENTIFICATION

Hamza Özer[1], *Bülent Sankur*[1], *Nasir Memon*[2]

[1]Department of Electrical and Electronics Engineering, Boğaziçi University, Bebek, Istanbul, TURKEY
2Department of Computer and Information Science, Polytechnic University, Brooklyn, NY, USA
hozer@uekae.tubitak.gov.tr, sankur@boun.edu.tr, memon@poly.edu

ABSTRACT

We propose and evaluate three new perceptual audio hash functions. These hash functions are based on a concise description of the time-frequency characteristics. This information can be extracted from the time series of frame-by-frame fundamental period or from the singular value decomposition of the mel frequency cepstral parameters. Experiments show that the proposed perceptual hash functions are both robust and unique.

1. INTRODUCTION

In this study we develop algorithms for summarizing a long audio signal into a concise signature sequence, which can then be used to identify the original record. We want this signature to be insensitive to non-malicious signal manipulations such as mild compression, but otherwise be sensitive to the content changes. This process is called robust audio hashing and the output sequence is denoted in the literature by alternate names, such as signature, fingerprint or perceptual hash values of the input. The mapping tool from audio input to the signature is called perceptual hash function.

Hash functions are deployed in the area of cryptology, where they are generally used to verifying the authenticity of data. In the cryptographic context, hash functions are required to be extremely fragile. In other words, any alteration of the source data, be it even one bit flipping, causes a totally different hash output. Instead, we are searching for robust hashes, which should resist against those signal-processing operations (filtering, compression or AD/DA conversion etc.) that purportedly leave the content intact.

Robust hashing finds applications in database searching, broadcast monitoring, tamper proofing, data content authentication etc. For example, in database searching and broadcast monitoring, instead of comparing the whole sample set, hash sequence would suffice to identify the content. Another application example is in watermarking, where a content-dependent signature, coupled with ownership or authorship label is embedded in the document. This type of watermarking is resistant, among other things, to copy attack.

The two desiderata of the perceptual hash function are robustness, and uniqueness. The uniqueness qualification implies that hash sequence is informative, that is, it should reflect the content of the audio document in a unique way. Such uniqueness is sometimes called randomness, so that any two distinct audio documents result in different and apparently random hash values. Consequently, the collision probability, that is the probability that two perceptually dissimilar inputs yield the same hash value, is minimized. The robustness qualification implies that the audio input can be subjected to certain non-malicious manipulations, such as analog-to-digital (A/D) conversion, compression, sample jitter, moderate clipping etc., and yet it should stay in principle the same in face of these modifications. The line of demarcation between what constitutes a non-malicious signal processing operation and when a change in content should be admitted depends upon the application.

There exists a number of perceptual audio hashing algorithms in the literature. The algorithms in [1,2] are exploiting the power spectrum of the signal and its statistical properties, and are intended, respectively, for audio database search and watermarking. The audio fingerprinting methods discussed in [3-6] are intended for music, speech and silence discrimination. They use principal component analysis (PCA), mel-frequency cepstral coefficients (MFCC), adaptive quantization and channel decoding to summarize the source data.

We investigate novel audio features for signature extraction and based on these, propose two perceptual audio hashing algorithms. One of them operates in the time domain, and uses the inherent periodicity of audio signals. The time profile of the dominant frequency of the audio track constitutes the discriminating information. The second one uses the time-frequency landscape, as given by the frame-by-frame MFCC coefficients and summarizes them via singular value decomposition.

2. PERIODICITY-BASED PERCEPTUAL HASHES

Our departure point is that all audio signals, be it speech, music or environmental sounds, have an inherent periodicity. The profile of the dominant period in the course of the audio record constitutes the signature of that signal. We employ two different approaches to measure the periodicity, which are estimation-based and correlation-based techniques, as shown in Fig. 1.

The input audio signal is split into smaller frames, which are in turn windowed by Hamming window to reduce the discontinuity effects. The number F of such frames is

This project was supported by AFRL Contract F30602-03-C-0091, NSF Project INT-021804 and TÜBİTAK project 102E018.

determined by the total length of the audio record and the frame size, N. The frames are further preprocessed in order to bring forward any periodicity that could be underlying the signal. Thus frame signals are effectively band-pass filtered by an LPC (linear predictive coding) filter to remove the short-term dependencies. Furthermore, the time series of estimated frame frequencies is smoothed as a postprocessing operation. This smoothing is enacted via a 7-tap moving average filter and is effective against desynchronization type of distortion.

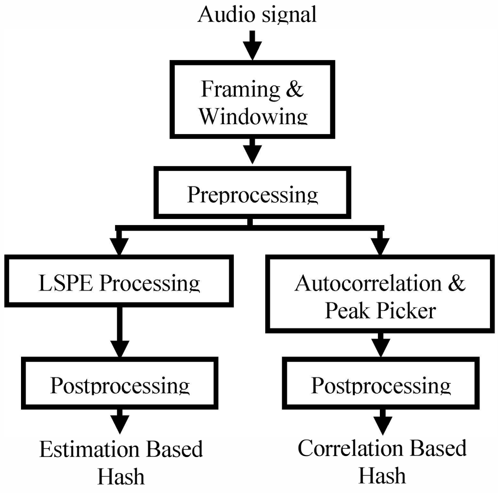

Figure 1. Block diagram of periodicity estimators.

2.1 Parametric Estimation of the Periodicity

A least-square periodicity estimator (LSPE) is applied to compute the period value of each frame. Irwin [8] had shown that applying a LSPE directly to the signal gives optimal periodicity detection. The LSPE calculation of a frame is as follows [9]: Let

$$s(i) = s_0(i) + n(i), \qquad for \quad i = 1,2,..., N$$

where *s(i)* is the input signal, $s_o(i)$ is the periodic component of input signal, *n(i)* the nonperiodic component of input signal, all within a frame of N samples. The periodic component satisfies the relationship $s_o(i) = s_o(i+kP_o)$ for integer k and for some period P_o.

We now let $\hat{P}_0$ be our estimate of P_0 and $\hat{s}_0(i)$ be the estimated periodic signal component, with period $\hat{P}_0$. An estimate $\hat{s}_0(i)$ is obtained from the input signal by:

$$\hat{s}_0(i) = \sum_{h=0}^{K_0} \frac{s(i+h\hat{P}_0)}{K_0}, \qquad 1 \le i \le \hat{P}_0 , \qquad P_{\min} \le \hat{P}_0 \le P_{\max}$$

where $P_{\min}$ and $P_{\max}$ delimit the range of P_0 and $K_0 = \left[{(N-i)}/{P_0} \right] + 1$ is the number of periods of $\hat{s}_0(i)$ in the analysis frame.

The objective of the least-squares method is to find the pitch period $\hat{P}_0$ that minimizes the mean square error $\sum_{i=1}^{N} [s(i) - \hat{s}_\bullet(i)]^2$ over each analysis frame. Friedman [9] shows that this is equivalent to maximizing $\sum_{i=1}^{N} \hat{s}_\bullet^2(i)$. However this estimate is biased towards large values of $\hat{P}_0$. To overcome this bias, Friedman derives the normalized periodicity measure:

$$R_1(\hat{P}_0) = \frac{I_0(\hat{P}_0) - I_1(\hat{P}_0)}{\sum_{i=1}^{N} s^2(i) - I_1(\hat{P}_0)}$$

where $$I_1(\hat{P}_0) = \sum_{i=1}^{\hat{P}_\bullet} \sum_{h=0}^{K_\bullet} \frac{s(i+h\hat{P}_0)^2}{K_0}$$ and

$$I_0(\hat{P}_0) = \sum_{i=1}^{\hat{P}_\bullet} \frac{\left[\sum_{h=0}^{K_\bullet} s(i+h\hat{P}_0) \right]^2}{K_0}.$$ For each frame, $R_1(\hat{P}_0)$ is computed for values of $\hat{P}_0$ between $P_{\min}$ and $P_{\max}$. The estimated period $\hat{P}_0$ is the argument that maximizes $R_1(\hat{P}_0)$.

2.2 Nonparametric Estimate of the Periodicity

A nonparametric periodicity estimate is obtained by peak picking in the correlation sequence of the preprocessed audio signal. In fact, the lag value of the first peak of the autocorrelation of the LPC filtered signal is used as a standard technique in speech analysis for pitch period estimation. Similarly in our work, we compute the period for each overlapped frame. The advantage of the correlation-based method is that it requires quite less computation then mean-square estimation method.

In the case where the audio signal does not possess an explicit periodicity, as in the case of unvoiced speech or silence, either estimation function generates a zero for that frame. Thus if the confidence in periodicity $R_1(\hat{P}_0)$ or the first correlation peak falls below 0.5, then we declare that frame as aperiodic.

3. SINGULAR VALUE DECOMPOSITION BASED METHOD

In this algorithm, we extract a signature from the audio signal by summarizing concisely the time course of its mel frequency cepstral coefficients (MFCCs) and thus we capture signal's time-frequency characteristics in the perceptual domain. The summarization is enacted by the Singular Value Decomposition (SVD) of the MFCCs of the signal, as organized in a matrix of features over frames.

It is known that mel frequency cepstral coefficients are useful short-term spectral-based features [7]. MFCC coefficients are calculated by first obtaining the short-time Fourier transform of the signal, then taking the logarithm of the magnitude spectrum, scaling and smoothing it over Mel-spaced frequency bins. The Mel-scale is based on a mapping between actual frequency and perceived pitch in accordance with the human auditory system's nonlinear perception. Finally, the MFCC features are obtained by applying some transform on the Mel-spectral vectors into time domain, for example, by the discrete cosine transform [7].

The MFCC features are organized in a *FxM* matrix form A, where each row consists of the MFCC values for a frame, and there are *F* rows, the number of frames into which the signal has been segmented. This matrix expresses the whole evolution landscape of the signal. A final summary of this landscape is computed by the SVD of the frame MFCC matrix.

The singular value decomposition, which is a factorization and summarization technique, effectively reduces the *FxM* MFCC-feature matrix into a much smaller invertible and square matrix. Thus the given *FxM* matrix is decomposed as $A = UDV^T$, where D is *FxM* matrix with only min(F,M) diagonal elements, and U is an *FxF* orthogonal matrix and V is an *MxM* orthogonal matrix. The diagonal entries of D are called the singular values of the matrix A, the column space of the matrix U are called the left singular vectors of the matrix A, and the column space of the matrix V are called the right singular vectors of the matrix A. In general, a few singular values (first few components of the diagonal of the matrix D) give a good summarization of the matrix A [10].

4. EXPERIMENTAL RESULTS

We have performed several experiments to evaluate on the one hand the robustness and on the other hand the uniqueness properties of the proposed perceptual hash functions. As audio database we used 3-5 seconds long utterances, which are sampled at 16kHz sampling rate. We conducted also some preliminary experiments with music files and obtained similar results.

Setting of the parameters: The setting of the feature parameters was as follows. For the periodicity estimator, for the 16kHz-sampled signal, $P_{\min}$ and $P_{\max}$ were set, respectively, to 40 and 320 samples, which means that the admissible periods are between 50Hz to 400Hz. The frames, taken to be 25 ms long, are overlapped by 50 percent. Frames are preprocessed by first low-pass filtering them with a cutoff frequency of 900 Hz and then through a 4-tap (Linear Predictive Coding) LPC inverse filter [7]. The resulting hash sequence consists of 79 samples/second.

For the SVD method, 13 cepstral features are obtained for each frame. Therefore an *Fx13* feature matrix is obtained from the input audio signal. We experimented with up to three singular values in the UD product, and it was observed that a single singular value was adequate. This is again the basic trade-off between uniqueness, which improves with more singular values, and robustness, which, conversely improves with smaller number of eigenvalues. The signature rate depends upon the number of frames and the number of singular values chosen, which becomes 26, 52 and 78 samples per second respectively, for the choice of 1 to 3 eigenvalues.

Types of attacks: We programmed eleven types of attacks to evaluate robustness performance. To this purpose, hash value of the original record is compared with the hash value of the attacked version. We measure their similarity with normalized correlation coefficient.

These attacks and their acronyms are as follows: 1) Comp: 3:1 compression below 10dB; 2) Subs: subsampling down to 8khz; 3) Ups: upsampling to 44.1 kHz; 4) NsyA: Noise addition, 20 dB SNR; 5) Dns1: Denoise filtering after noise addition; 6) Dns2: Denoise filtering of clear signal; 7) Pinc: Raise pitch 1%; 8) Pdec: Lower pitch %1; 9) Tcomp: time compress by %4; 10) Crp: random cropping, total amount %8; 11) TelF: telephone filtering, 135-3700Hz.

Normalized correlation is used as similarity measure between the hash sequence of the original sound file and that of the test file. Obviously this score takes values in the $(0,1)$ range.

Robustness tests: Table 1 summarizes the performance results of the three hash extraction methods, where EPM, CPM and SVDM refer to, respectively, estimation-based periodicity measure, correlation-based periodicity measure and singular value decomposition-based audio hashing methods. The resulting signature lengths are 79, 79 and 26 samples/second, in order, for the EPM, CPM and SVDM techniques. Notice that we could have made SVDM commensurate with dimension 78, but the performance difference between 1 and 3 eigenvalues was very small (less than 1%), so that we used the more parsimonious representation.

The EMP method performs slightly better then the CPM, albeit at a higher computational cost. However, SVDM produces the best results over considered attacks. The minimum similarities under all attacks for hash sequences are 0.80, 0.85 and 0.95 for the CPM, EPM, and SVDM methods, respectively.

Uniqueness tests: We tested whether the hash sequences can be confounded in a large repertoire of audio files. Thus, for each of the 200 utterances the hash value is computed and compared with all the other ones. The utterances are 3-4 seconds long distinct sentences, uttered by the same speaker. Notice that the use of only one speaker represents the worst case for confounding as we forego inter-speaker variability. Ideally, the similarity score between hashes should be zero. However, it has been observed that the maximum similarity (worst case) attained between hash values of different objects is 0.51, 0.56 and 0.70, respectively, for the EPM, CPM and SVDM methods. These experiments reveal that, in the case of uniqueness, SVDM is slightly inferior.

Table 1. Robustness tests against some unintentional types of attacks.

Method	NsyA	Dns1	Dns2	Pinc	Pdec	Comp	Subs	Ups	TelF	Crp	Tcomp
EPM	0.96	0.92	0.99	0.85	0.88	1	0.99	0.97	0.97	0.88	0.88
CPM	0.95	0.92	0.98	0.89	0.91	1	0.96	0.95	0.95	0.87	0.80
SVDM	0.99	0.99	0.99	.98	.98	.99	0.99	0.99	0.95	0.97	0.99

The correlation scores are given in Fig. 2. The dispersion of the right histogram shows the degree to which the hash value is affected by the signal processing attacks. The left histogram indicates the randomness of the hashes. In fact hash values also depend upon the content. It can be stated that, similarity between hash values of original object and its distorted ones are well separated from that of distinct audio records.

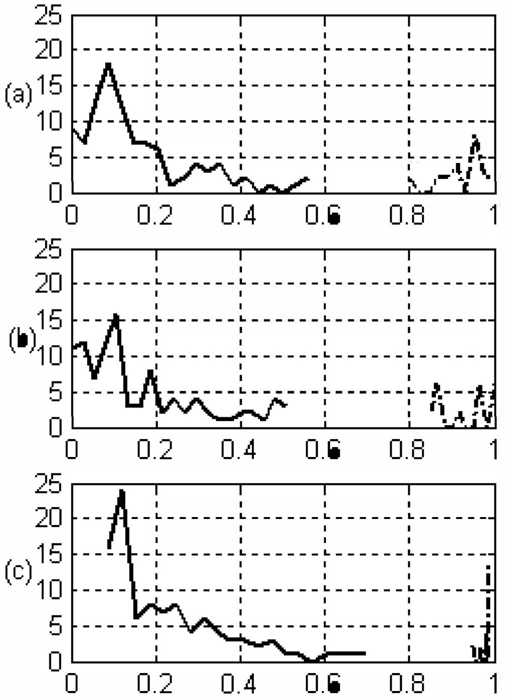

Figure 2. Histograms of the similarity measures of the hash values extracted from distinct objects (solid lines), and extracted from distorted versions of the same object (dashed lines). The abscissa plots the similarity score, while the ordinate shows the histogram value: (a) EPM, (b) CPM, (c) SVDM.

More specifically, the minimum similarity between attacked versions of the same object is always higher than the maximum similarity between distinct objects for all measures. In fact, this gap measures 0.85-0.51 for EPM, 0.80-0.56 for CPM and 0.95-0.70 for SVDM.

5. CONCLUSIONS AND FUTURE WORK

The three proposed perceptual audio hashing methods were tested, on the one hand, to measure for robustness vis-à-vis non-malicious signal processing attacks, and on the other hand, to assess the uniqueness or randomness of the hash when audio files with different content. All three proposed perceptual hash sequences perform satisfactorily, with SVDM being more robust. Our study continues in the direction of quantization or/and an encoding schemes for the hash sequences.

REFERENCES

[1] T. Kalker, J. Haitsma, and J. Oostveen, “Robust audio hashing for content identification”, *Int. Workshop on Content Based Multimedia Indexing,* Brescia, Italy, September 19-21, 2001.

[2] M. K. Mıhçak and R. Venkatesan, “A perceptual audio hashing algorithm: a tool for robust audio identification and information hiding”, *Inf. Hiding* 2001, 51-65.

[3] C.J. Burges, J. C. Patt, and S. Jana, “Distortion discriminant analysis for audio fingerprinting”, *IEEE Transaction on Speech and Audio Proc.,* Vol 11, No:3, pp. 165-174, 2003.

[4] F. Kurth and R. Scherzer, “Robust real-time identification of PCM audio sources”, *Presented at 114th Convention of Audio Engineering Society,* Amsterdam, The Netherlands, March 22-25, 2003.

[5] S. Sukittanon and L. E. Atlas, “Modulation frequency features for audio fingerprinting”, *in Proceedings of the 2002 IEEE ICASSP*, 2002.

[6] Beth Logan, “Mel frequency cepstral coefficients for music modeling”, in ISMIR, October, 2000.

[7] L. R. Rabiner and R. W. Schafer, Digital Processing of Speech Signals, Prentice-Hall, 1978.

[8] M.J. Irwin, “Periodicity estimation in the presence of noise”, *Inst. Acoust. Conf. ’79,* Windemere, UK.

[9] D.H. Friedman, “Pseudo-maximum-likelihood speech pitch extraction”, *IEEE Trans. ASSP-25,* (3), pp. 213-221, 1978.

[10] D. Wu, D. Agrawal, A. E. Abbadi, “Efficient retrieval for browsing large image databases”, *Proc. of the 5th Int. Conf. on Knowledge Management,* pp. 11-18, Rockville, MD, November, 1996.